REFERENCE

Transcribing across languages

Which languages are hard, how code-switching breaks recognition, and a workflow for multilingual audio.

A 2026 reference for multilingual audio

For people who work with audio in more than one language

2026 edition

ABOUT THIS REFERENCE

About this reference

Most writing about AI transcription quietly assumes English. This reference is for everyone whose audio is not: journalists working in two languages, researchers interviewing across borders, global teams whose meetings drift between tongues mid-sentence. It summarizes where multilingual speech-to-text stands in 2026, drawing on published benchmarks and vendor documentation from the first half of the year.

Prepared as an open reference. Share freely.

1. The short version

For a handful of well-resourced languages, recognition is close to solved. For everything else, quality drops along a curve that tracks how much training data existed, not how "hard" the language sounds. The real failure mode in multilingual work is not a difficult accent, it is switching languages inside one recording. That breaks more tools than any single language does.

2. Why some languages are harder

Speech recognition quality is mostly a data problem. A model trained on hundreds of thousands of hours of English will read English well and read a low-resource language poorly, regardless of how the two languages compare linguistically. The result is a rough tiering.

Tier	Examples	What to expect
High resource	English, Spanish, French, German, Mandarin	Near-parity on clean audio, small gaps between tools
Medium resource	Portuguese, Dutch, Arabic (MSA), Korean, Italian	Good on clean audio, more errors on accent and slang
Lower resource	Regional dialects, smaller national languages	Usable draft at best, plan for real correction

Two extra factors sit on top of raw data volume. Languages with heavy dialect variation, such as Arabic, perform far worse on colloquial speech than on their standard written form. And morphologically rich languages, where one word carries many inflected endings, give recognizers more ways to be slightly wrong.

3. Code-switching is the real problem

The single most underestimated issue in multilingual transcription is code-switching: a speaker starting a sentence in one language and finishing in another. It is normal in bilingual workplaces, in technical fields where English terms pepper the local language, and across much of the multilingual world.

It breaks recognition because a model predicts the next word partly from the language it thinks it is hearing. When the language changes mid-clause, that assumption is wrong for a moment, and errors cluster right at the switch. A tool can score well on monolingual benchmarks and still fall apart on a real bilingual meeting. The only reliable test is your own mixed-language audio, not a clean sample in a single language.

4. Accuracy by language, honestly

The advertised figure, usually 95 to 99 percent, describes clean audio in a high-resource language with one speaker. Under real conditions the number moves.

Condition	High-resource language	Medium-resource language
Clean, single speaker	95 to 98 percent	90 to 95 percent
Accent or dialect	90 to 95 percent	85 to 90 percent
Code-switching	drops noticeably	drops sharply
Noise or crosstalk	degrades for both	degrades for both

> The practical reading: treat any published accuracy number as a ceiling for the best case, then discount it for your language tier, your speakers, and your recording conditions.

5. A workflow for multilingual teams

A few habits that save time when audio spans languages.

Pick one tool with broad language coverage rather than a separate specialist per language. Juggling several single-language tools costs more time in format-wrangling than the marginal accuracy is worth for most work.

Transcribe in the source language first, then translate as a second step if you need it. Translating during recognition compounds two error sources into one messy output; separating them keeps each step checkable.

Budget correction time per language, not once. A pass that takes five minutes for English audio may take fifteen for a lower-resource language. Plan the human minute where the tool is weakest.

Keep proper nouns in a small glossary. Names, places, and product terms are the most common errors across every language, and a quick find-and-replace list handles most of them.

6. A representative example

To ground the abstractions: a consumer tool like Vomo.ai shows the multilingual shape in one product. It transcribes across 50-plus languages, stating 95%+ accuracy on clear audio, adds punctuation, and separates speakers. On top of the transcript it builds a summary, key points, and action items, and lets you query the recording in plain language with answers drawn from the actual text. Input can be a live recording, an uploaded file, or a pasted video link, and output exports as TXT, DOCX, PDF, SRT, or Markdown. It runs on the web and iOS with a free tier for evaluation.

Naming it is not a ranking. It is a clean example of what broad language coverage plus a comprehension layer looks like in practice.

More at https://vomo.ai/

7. Common questions

Which languages are safe to trust with light review? The high-resource tier on clean audio: major European languages and Mandarin, with a single clear speaker. Everything else deserves a closer read.

How do I handle a meeting in two languages? Test candidates on a real bilingual recording before committing. Expect errors to cluster at the language switches, and plan correction time there specifically.

Is local, offline transcription an option for privacy? Yes. Open models such as Whisper can run locally and support many languages, at the cost of setup effort and compute. For sensitive multilingual material it is worth the trouble.

This reference may be shared and quoted with attribution. Figures reflect published benchmarks and vendor documentation available in the first half of 2026 and should be re-checked before use in procurement decisions.